Algorithm for Minimizing Rebate Value due to SLA Breach in a Utility Computing Environment

DESCRIPTION

Cross-Reference to Related Application

[Para 1] The present invention is related to the subject matter of U.S. Patent Application _____ (Attorney Docket number AUS920030302US1), and U.S. Patent Application _____ (Attorney Docket No. RSW920030148US1), both of which are incorporated herein by reference.

Field of the Invention

[Para 2] This invention generally relates to the field of managing resources in a utility computing environment. In particular, this invention relates to a process for minimizing the total value of rebates disbursed to customers, in a utility computing environment, as a result of a service level agreement breach.

Background of the Invention

[Para 3] For years, information technology (IT) organizations (the "providers") have offered IT management services and computing resources to other businesses (the "customers") within a utility computing environment. While a customer may purchase or lease IT resources directly from a provider for that customer's exclusive benefit, a customer also may share a provider's computing resources and management services with other customers. In a typical utility computing environment, the customer does not purchase or lease the physical resources; instead, the provider retains the discretion to allocate the resources as needed to meet its service obligations. Nonetheless, the provider must meet the requirements of each customer sharing the IT resources as specified in a contract or an agreement. If these service requirements are not met, the provider has breached its service obligation to the customer and the provider must compensate the customer for the breach.

[Para 4] As illustrated in FIG. 1, customers of on-demand services share management and computing resources (to the system and subsystem level), including persistent memory ("storage"), volatile memory ("memory"), and processors. FIG. 1 portrays another characteristic of the on-demand model - multiple customers sharing the same subsystem within the same computing resource, such as a logical partition (LPAR). In FIG. 1, for example, customer **3** and customer **4** each could run separate instances of operating system **3**, such as International Business Machines, Inc.'s (IBM) Z/LINUX, on a single Z/VM (also by IBM) LPAR. When multiple external customers share the same hardware, as described here, performance tuning of the system must be applicable to both the workload and to all customers sharing the hardware.

[Para 5] A Service Level Agreement (SLA) typically is used in an on-demand shared environment to establish to threshold levels of service and guide the dynamic allocation of IT resources. The SLA is a contract, or series of contracts, that embodies the mutual understandings between the provider and the customer. Thus, any failure to provide the agreed level of service to a customer is referred to herein as an "SLA breach" or "breach." The SLA also sets system (and subsystem) performance expectations and defines the procedures and reports needed to track compliance to the agreement. The SLA may contain the process for reporting service problems, the time frame for problem resolution, the process for monitoring service levels, and the penalties associated with any given SLA breach.

[Para 6] A performance monitoring tool, commonly referred to as a profiling tool, collects performance data to determine compliance with the SLA. The profiling tool tracks and measures performance characteristics of the system including CPU utilization, processing time, and the memory or storage available to a customer. Often, these tools are designed to operate in a particular environment. Performance Monitoring Infrastructure Request Metrics is an example of a profiling tool designed to operate after deployment in a web-based environment. See, generally, http://publib.boulder.ibm.com/infocenter/ws51help/index.jsp?topic=/com.ibm.websphere.exp.doc/info/exp/ae/tprf_requestmetrics.html. Additionally, system administrators use the information obtained from these performance measurements ("metrics") to tune the performance of the system and take corrective action if needed. When the profiling tool indicates that system resources are not available, or are not performing according to the SLA, the SLA is breached. The provider pays a penalty to compensate the customer for the SLA breach according to the terms of the SLA.

[Para 7] One of the fundamental tenets of a utility computing environment is the concept of proactively rebating, i.e., compensating, a customer when an SLA is

breached. In a typical on-demand scenario, the various customers hosted by a single provider agree to different levels of service and compensation or "rebate" for an associated breach. For example, some of these customers may be "premium" customers, who pay more for higher service levels and are entitled to greater compensation when there is an SLA breach. These premium customers consequently represent a greater penalty to the provider in the event of an SLA breach. Other customers may subscribe as "standard" customers, who pay relatively less for the services, receive less compensation when there is a breach, and thus, represent a lesser degree of penalty in the event of an SLA breach. A sample scenario is provided in FIG. 2.

[Para 8] In addition to using profiling tools, there are several methods available to IT service providers in the utility computing environment to measure compliance with an SLA. Some of these methods also calculate appropriate rebates to customers in the event of an SLA breach, and proactively disburse a rebate to a customer. These processes are disclosed in U.S. Patent No. 6,195,697 (issued February 27, 2001), U.S. Patent No. 6,556,659 (issued April 29, 2003), and U.S. Patent Application No._____ (Attorney Docket No. AUS920030302US1). These processes do not address optimizing network resources and managing conflicting needs among the customers of the shared network collectively, nor do these processes address reallocating resources among the customers to minimize the total rebate awarded in the event of an SLA breach.

[Para 9] Patent App. No. 0062205 (published April 1, 2004) assigns a financial value to identified performance flows based on SLA requirements and penalties for breach of the requirements. This financial value alerts operators of the possible financial impacts of reconfiguring hardware or software associated with those identified flows. This process, however, merely calculates and displays the financial loss associated with a breach or potential breach of one individual customer's SLA. U.S. Patent Application No. _____(Attorney Docket No. RSW920030148US1) does provide a method for estimating an SLA breach value, based on data acquired from an individual customer and on data acquired from an aggregated group of customers. But again, this method does not disclose a means for minimizing the total rebate a service provider must offer when an SLA is breached.

[Para 10] Thus, the tools used to track and measure the performance characteristics of transactions throughout a system to determine compliance to an SLA are common. Similarly, processes for calculating the rebate that a service provider must proactively award to a customer when the SLA is breached are not new. There is not, however, a

tool or process available to service providers for minimizing the total rebate a service provider awards in the event of an SLA breach.

[Para 11] Rebates in the form of monetary compensation, free software, or other forms, are costly to service providers. Rebates affect a provider's overall profitability as well impact the provider's goodwill. After all, those customers who have paid a premium price for service are not receiving the level of service agreed upon. These customers may suffer financial losses and losses of goodwill, as well, if they, in turn, cannot meet their business demands. Therefore, one skilled in the art should appreciate the advantages of an invention that precisely addresses the problem of minimizing rebates the service providers disburse to customers as a result of an SLA breach. This and other objects of the invention will be apparent to those skilled in the art from the following detailed description of a preferred embodiment of the invention.

Summary of the invention

[Para 12] The invention described is a new and useful process for minimizing the overall rebate a provider disburses when an SLA breach occurs in a utility computing environment. The inventive process calculates the total minimum rebate value payable by a provider to a customer, or group of customers, in the event of an SLA breach. The process compares performance data and resource usage with the SLAs of the customers, and reallocates shared resources to those breached customers who represent a lesser penalty to the provider in the event of an SLA breach. Specifically, the process determines if there is a breach, and if so, identifies the breached customer and the breached customer's status based on the penalty provided in the SLA. The process also identifies the underlying resource causing the breach. The process then creates a list of customers with a lower customer status and determines if any of the customers with the lower status are under-utilizing the resource. If the resource is under-utilized, the process then reallocates these under-utilized resources to those breached customers requiring additional resources to meet SLA thresholds. If all resources are operating at peak capacity, the process reallocates the resources to those customers whose SLAs provide a greater penalty in the event of an SLA breach, as compared to those customers whose SLAs provide for a lesser penalty, thereby minimizing the total rebate due upon an SLA breach.

Brief Description of Drawings

[Para 13] The novel features believed characteristic of the invention are set forth in the appended claims. The invention itself, however, as well as a preferred mode of use, further objectives and advantages thereof, will be understood best by reference to the following detailed description of an illustrative embodiment when read in conjunction with the accompanying drawings, wherein:

[Para 14] FIG. 1 illustrates an exemplary shared resource configuration;

[Para 15] FIG. 2 depicts an exemplary customer scenario in a utility computing environment;

[Para 16] FIG. 3 represents an exemplary prior art network of computers and other hardware devices, in which the Rebate Minimization Algorithm may be implemented;

[Para 17] FIG. 4 is a schematic diagram of the host server memory used to implement the Rebate Minimization Algorithm; and

[Para 18] FIG. 5 depicts the inventive process for minimizing rebates disbursed to customers sharing IT resources.

Detailed Description of the Preferred Embodiment

[Para 19] The present invention is a process for minimizing the total rebate value that an IT provider disburses to customers in the event of an SLA breach. The invention, although operable in a variety of hardware and software configurations, operates in a utility computing environment wherein an IT service provider allocates shared IT resources to customers according to the terms of previously established SLAs.

[Para 20] The term "computer hardware" or "hardware," as used herein, refers to any machine or apparatus that is capable of accepting, performing logic operations on, storing, or displaying data, and includes without limitation processors and memory; the term "computer software" or "software," refers to any set of instructions operable to cause computer hardware to perform an operation. A "computer," as that term is used herein, includes without limitation any useful combination of hardware and software. A "computer program" or "program" includes without limitation any software

operable to cause computer hardware to accept, perform logic operations on, store, or display data. A computer program may, and often is, comprised of a plurality of smaller programming units, including without limitation subroutines, modules, functions, methods, and procedures. Thus, the functions of the present invention may be distributed among a plurality of computers and computer programs. The invention is described best, though, as a single computer program that configures and enables one or more general-purpose computers to implement the novel aspects of the invention. For illustrative purposes, the inventive computer program will be referred to as the "Rebate Minimization Algorithm" (RMA).

[Para 21] The RMA calculates the total minimum rebate value payable by an IT service provider to a customer, or group of customers, in the event of an SLA breach. As used herein, the term "service provider" or "provider" refers to any entity that provides management services and computing resources to any individual or entity. A "customer" is any individual or entity acquiring the management services and shared computing resources from the service provider.

[Para 22] Notably, the RMA determines if the provider has breached an SLA. If so, the RMA reallocates under-utilized resources to meet the demands of the breached customer if the breached customer represents a greater penalty to the provider than those customers under-utilizing the resources. A "penalty," as that term is used here, refers to the amount of compensation or rebate a provider must pay a customer for breaching an SLA. The compensation the customer receives from the provider determines the customer's "status." If all resources are operating at peak capacity, the RMA reallocates the resources used by customers with a lower customer status to the customers with a higher customer status, thereby minimizing the total rebate cost associated with an SLA breach.

[Para 23] As noted above, in a utility computing environment, a service provider offers management services and computing resources to a customer at the system and subsystem level. Inasmuch as the customer acquires services and resources from the provider, the customer may, in turn, offer goods, services, or information, for purchase, lease, or use to other individuals or entities, usually via the Internet. Any such individual or entity purchasing, leasing, or otherwise obtaining or using goods, services, or information from the customer is referred to herein as a "consumer." In other words, the consumer may purchase items on-line from the customer's website. The consumer communicates with the customer by means of a network, routed through a utility computing environment, which the provider maintains. The amount of IT resources available to the customers, in the utility computing environment, determines the number of consumers a customer may host on its website at any

particular instance, and thus, further determines the amount of business that a customer may transact.

[Para 24] The inventive RMA is described in detail below with reference to an exemplary prior art network of hardware devices, as depicted in FIG. 3. A "network" comprises any number of hardware devices coupled to and in communication with each other through a communications medium, such as the Internet. A "communications medium" includes without limitation any physical, optical, electromagnetic, or other medium through which hardware or software can transmit data.

[Para 25] For descriptive purposes, exemplary prior art network **100** has a limited number of nodes, including consumer workstation computer **105**, consumer workstation computer **110**, consumer workstation computer **115** (collectively consumer workstation computers **105-115**), host server computer **120**, database server computer **125**, and database **130**. The term "server" refers to a computer system that is shared by multiple clients. A server may refer to the entire computer system, i.e., hardware and software, or just the software that performs the service. For example, the term "database server", as used herein, refers to the both the hardware and software necessary to store and retrieve data. In contrast, the term "web application server", as used herein, refers to any software product designed to operate in a web-environment, such as an HTTP server that manages requests from a browser and delivers HTML documents and files in response. Web server software is frequently used in e-commerce and executes server-side scripts, such as Java Script and Java server pages (JSPs), to retrieve data from a database and display the data in the form of a web page via browsers or client applications. The term "host server computer" refers to the hardware on which the RMA and customers' resources reside. A person of skill in the art also should appreciate that a database may exist in many forms. As used herein, the term "database" generally refers to any collection of data stored together and organized for rapid search and retrieval, including without limitation flat file databases, fielded databases, full-text databases, object-oriented databases, and relational databases. While host server computer **120**, database server computer **125**, and database **130** are further located within utility computing environment **135**, consumer workstation computers **105-115** are outside of the utility computing environment **135**.

[Para 26] Host server computer **120** hosts programs, applications, and tools that control consumption of computing resources in utility computing environment **135**. Therefore, in this embodiment, the consumer accesses the utility computing environment **135** via consumer computer workstations **105-115** networked to host

server computer **120** by network connection **140**. The amount of IT resources available to the customer, as allocated by applications installed on host server computer **120**, determine the actual number of consumers that may access the customer's website, and thus determine the number of consumers that a customer may service. Network connection **140** comprises all hardware, software, and communications media necessary to enable communication between network nodes **100-130**. Consumer workstation computers **105-115** use publicly available protocols or messaging services to communicate with the host server computer **120** through network connection **140**. Host server computer **120** interacts with database server **125** to store and retrieve SLA information **145**, system performance data **150**, and customer information **155** to and from database **130**. SLA information includes the level of service on which the customers and providers have agreed, and the penalty the provider pays when the provider breaches the SLA. Performance data, obtained in a web application environment, may include the retrieval rate from the web application server to the database, the time to request and return a displayed web-page, and the number of pooling requests, i.e., the number of users, an application can manage. Customer data may include information regarding inventory, shipping, prices, and consumer records. The nodes in the utility computing environment **135** also use publicly available network protocols; however, a firewall may control access to the utility computing environment **135**.

[Para 27] Memory **200** of host server computer **120** typically contains various applications such as web application server **205**, profiling tool **210**, and RMA **215**, as depicted in FIG. 4. The term "memory," as used herein, includes without limitation any volatile or persistent medium, such as an electrical circuit, magnetic disk, or optical disk, in which a computer can store data or software for any duration. A single memory may encompass and be distributed across a plurality of media. Memory **200** may include additional data and applications. Memory **200** also contains customer 1 memory resource **220**, customer 2 memory resource **225**, and customer 3 memory resource **230** (collectively customer memory resources **220-230**), as represented schematically in FIG. 4. Web application server **205** executes server-side scripts, such as Java Script and JSPs, to retrieve data from a database and transmit data in the form of a web page to the consumer workstation computers **105-115**.

[Para 28] Profiling tool **210** collects performance data **150** by tracking and timing individual transactions within utility computing environment **135**. Web application server **205** actively allocates customer memory resources **220-230** as part of its normal operation. FIG. 4 is included as a descriptive expedient and does not necessarily reflect any physical embodiment of memory **200**. Notably, customer memory resources **220-230** represent any number of shared resources. The term "shared resource" includes any computing resource that the service provider allocates

among various customers according to the terms of the customer's SLAs. Although volatile memory is depicted in FIG. 4 as the allocated resource, other resources, such as persistent memory, CPU utilization, and network bandwidth may be provisioned according to the SLA. For descriptive purposes, the applications are stored on host server computer **120**, but these applications may be located on any server which the host server computer is capable of accessing.

[Para 29] As FIG. 5 depicts, RMA **215** determines if there is an SLA breach or a potential SLA breach, by comparing the resource usage of the customers, as indicated by performance data **150**, with the corresponding customer SLA information **145** (**410**). If there is an actual or potential breach, RMA **215** identifies the breached customer, the breached customer's status, and the underlying resource causing the breach (**420**, **430**, and **440**). Customer status is determined by the severity of the penalty, i.e., the amount the provider rebates the customer, for services not rendered, when the provider breaches the SLA. RMA **215** then determines if there are any customers, sharing the underlying resource causing the breach, with a lower customer status than the breached customer (**450**). If there are no customers with a lower customer status than the breached customer, the current breach scenario is the optimal scenario, no resources are allocated, and the process ends (**510**). If there are customers sharing the underlying resource causing the breach, then RMA **215** next creates a target customer list, which includes all customers whose status is lower than the breached customer's status (**460**). RMA **215** determines if the resource usage of customers included in the target customer list is less than specified in the corresponding SLAs, i.e., the resources are not operating at peak capacity and therefore are under-utilized (**470**). If the resources are under-utilized, RMA **215** reallocates the under-utilized resources to the breached customer to minimize the rebate value disbursed for the SLA breach, and the process ends (**480** and **510**). In contrast, if RMA **215** determines that all customers are operating at peak capacity by using the resources to the maximum extent specified in the SLA so that the resource is not under-utilized, RMA **215** releases the resource allocated to customers having a lower customer status, as indicated in the target customer list. RMA **215** then reallocates the resources to the breached customer who represents a greater penalty to the provider (**470**, **490**, and **500**). RMA **215** thus determines the minimum total rebate payable by the provider for an SLA breach and the process ends (**510**).

[Para 30] A preferred form of the invention has been shown in the drawings and described above, but variations in the preferred form will be apparent to those skilled in the art. The preceding description is for illustration purposes only, and the invention should not be construed as limited to the specific form shown and described. The scope of the invention should be limited only by the language of the following claims.